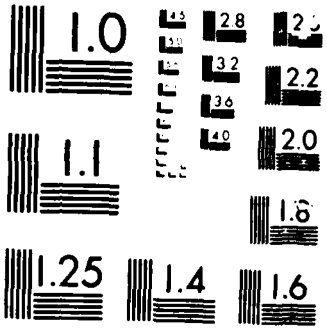
AD-A176 475 DISTRIBUTED SYSTEM MODELLING AND ANALYSIS(U) 1/1
MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR COMPUTER
SCIENCE N A LYNCH 22 DEC 86 ARO-20519.35-EL
UNCLASSIFIED DAAG29-84-K-0058 F/G 9/2 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0

2

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ARO 20519.35-EL

2. GOVT ACCESSION NO.
N/A

3. RECIPIENT'S CATALOG NUMBER
N/A

4. TITLE (and Subtitle)
"Distributed System Modelling and Analysis

5. TYPE OF REPORT & PERIOD COVERED
Final report; April 1, 1984 - Sept. 30, 1986

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Nancy A. Lynch

8. CONTRACT OR GRANT NUMBER(s)
DAAG29-84-K-0058

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Massachusetts Institute of Technology
Laboratory for Computer Science
Cambridge, MA. 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
December 22, 1986

13. NUMBER OF PAGES
21

MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Distributed algorithms, lower bounds, distributed consensus, Byzantine agreement concurrency control, nested transactions.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This project has made considerable progress in developing theoretical foundations for distributed computing. The primary thrust of the work has been the design of distributed algorithms and the proof of upper and lower complexity bounds for distributed problems. The kinds of problems studied include distributed consensus in the presence of faults, resource allocation, and election of a leader. A secondary effort has involved the development of formal semantic models for distributed algorithm. A tertiary effort has involved the modelling, specification and verification of concurrency control and recovery algorithms for nested

ARo 20519.35-EL

# "Distributed System Modelling and Analysis"

Final Report

Professor Nancy A. Lynch

December 23, 1986

U.S. ARMY RESEARCH OFFICE

Contract Number DAAG29-84-K-0058

Massachusetts Institute of Technology

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

THE VIEW, OPINION, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

## TABLE OF CONTENTS

## a. Statement of the Problem Studied

This project is intended to develop theoretical foundations for distributed computing. The primary goal of the work has been the design of distributed algorithms and the proof of upper and lower complexity bounds for interesting distributed problems. The kinds of problems studied include distributed consensus in the presence of faults, resource allocation, and election of a leader.

A secondary goal has been the development of formal semantic models for concurrent and distributed algorithms, in a way which would clarify the commonality among various different kinds of concurrent algorithms (shared memory algorithms, message-passing algorithms, concurrency control algorithms, dataflow algorithms, etc.)

A tertiary goal has involved the modelling, specification and verification of concurrency control and recovery algorithms for nested transaction systems.

## b. Results

### I. Analysis of Algorithms

### A. Distributed Consensus

In [DLS], we devise algorithms for the problem of reaching agreement in a realistic distributed system model that lies between the completely synchronous and completely asynchronous models. In this model, messages have a "usual delivery bound", which need not always hold. In our solutions, disagreement can never be reached, no matter how the messages behave. Moreover, if the messages get delivered within their usual delivery bound for a sufficiently long interval of time, then agreement is guaranteed. Algorithms are given for various fault models, along with matching lower bounds.

The first version of [DLS] included separate proofs for all the results. In preparing a journal version of [DLS], we discovered a better way of organizing the results. Namely, we discovered a natural abstract partially synchronous model which can be used to present the algorithms, and a collection of reductions which allow the various models in the paper to simulate the abstract model.

In [BL,CDDS], we introduce and study a new and fundamental problem for distributed systems, which we call the Byzantine Firing Squad Problem. The problem is for remote processes to manage to carry out some specified action at the same time, in a setting where the processes wake up at different times, and where some of the processes are faulty. We obtain algorithms and lower bounds for a variety of fault models. Our bounds are tight for all but one of the models.

In [FLM], we demonstrate a new technique for proving lower bounds on the number of processors needed to solve various distributed consensus problems. We have been able to unify a large collection of previous work on impossibility for various kinds of distributed consensus, and add several new results, using a new "shifting scenarios" technique. Many of the results were previously known, with very complex proofs. There are some new results, however, in the area of clock synchronization. The paper was the highest-rated submission to the 1985 PODC conference, and was invited to appear in the flagship issue of the new Springer-Verlag Journal on Distributed Computing.

The work in [CMS] provides lower bounds for the expected time to reach Byzantine agreement in a variety of fault models.

In [CC], we have developed a randomized Byzantine agreement algorithm that terminates in an expected number of rounds that is smaller than the known lower bound (due to Lynch and Fischer) on the number of rounds required by a deterministic Byzantine agreement algorithm. The algorithm is of interest for several reasons. It is simple and efficient enough to be of practical importance. Also, it is an example of a situation where randomization improves on the problem solving power of a system of computers.

Although randomization is vital to the algorithm, it is used sparingly: the expected number of coin tosses per processor is less than one.

Brian Coan [C1] has developed a two-step transformation of algorithms in various fault models (fail-stop, failure-by-omission, and Byzantine), to a communication-efficient normal form. The first step is a transformation into a well known communication-inefficient normal form in which each processor, at each round, broadcasts its entire state. The second step is a new transformation from this communication-inefficient form to to a communication-efficient normal form. For each fault model there is a separate transformation. The transformation in the Byzantine model is fully worked out, and more work is needed for the other fault models.

As a corollary to the results in the Byzantine fault model, Brian obtained a major new result about the communication requirements of Byzantine agreement. This new result is a polynomial-message Byzantine agreement algorithm that uses about half the rounds of communication used by any other polynomial-message algorithm.

The problem of achieving simultaneity in the presence of faults first appeared implicitly in work of Rabin. He had an algorithm for reaching consensus whose expected running time was constant; however, different processors might terminate at different rounds. The implicit question was: "Does there exist an algorithm for achieving simultaneity that runs in time strictly less than O(t) (the lower bound for agreement in a deterministic algorithm)?". We have answered this question negatively, in [CD]: we have shown that not only is there no fast deterministic algorithm for achieving simultaneity, but there is not even a randomized algorithm whose expected running time is less than $t+1$ rounds, where the expectation is taken over the coin flip sequences. These results only assume fail-stop faults, and therefore apply *a fortiori* to more malicious failure models.

This work was continued by Dwork with Yoram Moses [DM]. They weakened the restrictions in [CD], on the failure patterns for which the lower bound could be proved. In fact, they have actually been able to completely characterize the time requirements (at least for certain consensus problems in which simultaneous termination is required, and for stopping faults). That is, they are able to exhibit a simple protocol that is optimal in the sense that it always halts at the earliest possible time, given the pattern in which the processors fail. This is often much earlier than the best previously known protocol for this problem.

This work was further continued by Yoram Moses and Mark Tuttle [MT]. In this continuation, they apply the theory of knowledge in distributed systems to important problems in distributed computing. Whereas the paper [DM] analyzes simultaneous Byzantine agreement in the crash failure model by studying when facts become common knowledge, the present paper extend the previous paper to general simultaneous actions in a variety of models of omission failures. These papers show that the major issue in designing protocols for simultaneous actions in unreliable systems is the uncertainly that individual processors have about other processors' views of the system. These papers demonstrate that a knowledge-based analysis can provide substantial insight and improved protocols for such problems. In particular, they show that it is possible to design protocols for simultaneous actions that in all of their runs will halt at the earliest possible time, given the behavior of the system.

Coan and Lundelius [CL] have studied the transaction commit problem in a realistic partially synchronous computation model. Namely, they assume that message delays and relative processor speeds are unbounded, and processors are subject to stopping faults. The time behavior of the system during an execution influences the correctness conditions as follows: if any processor votes to abort, then all processors must decide to abort; if all processors vote to commit and if no processors fail and all messages arrive within a known time bound, then all processors must decide to commit. The nonfaulty processors must always agree on their decision. In this model, they describe a randomized transaction commit protocol based on Ben-Or's randomized asynchronous Byzantine agreement protocol. The expected number of asynchronous rounds until the protocol terminates is a small constant, and the number of stopping faults tolerated is optimal. It is known that no deterministic protocol is possible in this model.

### B. Approximate Agreement

In [DLPSW], we give a new algorithm and matching lower bound for the problem of reaching approximate agreement (for example, agreement on the value of a sensor) among processors in a distributed network. Interestingly, the problem turns out to be considerably easier than the problem of reaching exact agreement. In particular, our solutions work in asynchronous networks with faults, whereas it has been previously shown that no solution to exact agreement is possible in such a network.

A version of [DLPSW] was prepared, submitted and accepted to JACM. New results were obtained, showing how only 3t+1 processors suffice to reach approximate agreement in an asynchronous environment with t faults, and showing how faulty processors can be rendered unable to determine the worst-case running time for the algorithm. A new lower bound was also obtained for the rate at which the approximation can converge.

Alan Fekete has obtained some preliminary results which show how the theoretical bound on rate of convergence can be approached by an actual algorithm.

### C. Clock Synchronization

In [LuL1,LuL2,Lu1], we study the problem of synchronizing software clocks in a distributed system. [LuL1] contains a new clock synchronization algorithm for use in a system in which some processors exhibit worst-case (Byzantine) faults; it is able not only to maintain synchronization, but also to bring the clocks into synchronization in the first place. Moreover, it enables easy recovery of failed processors. [LuL2] contains a surprising lower bound on the closeness with which clocks can be synchronized in the presence of uncertainty in the message delivery time. The lower bound is shown to be tight.

Jennifer Lundelius implemented a slightly modified version of the clock synchronization algorithm from [LuL1] at AT&T Bell Laboratories this summer. The program was written in the C language and was designed to synchronize the clocks of Suns running Berkeley Unix on an Ethernet. The algorithm had to be modified in an interesting way because of the reality of the Ethernet -- it does not provide reliable, bounded delay communication as well as a broadcast primitive. This paper describes the necessary modifications, analyzes the worst-case performance of the new algorithm, and gives an overview of the program.

### D. Electing a Leader

In [FL], we study the communication cost of the very important problem of electing a leader in a network of processors. Our results are for the special but important case of a synchronous, bidirectional ring network. They show that any algorithm which solves this problem must use at least order n log n messages. An interesting combinatorial technique is used.

### E. Network Resource Allocation

[FGGL,LGFG,FLBB] are papers about network resource allocation. Most of the results in these papers were obtained a couple of years ago; however, we have been settling some open cases and polishing up the presentation.

[LGFG] was finally completed and sent to Information and Control. Over the past year or more, we have improved this work in many ways. The most recent improvements involve generalizing the analysis to allow arbitrary probability distributions of request arrivals, and to the case where resources, as well as requests, occur at locations that are determined probabilistically.

[LGFG] contains criteria for optimal placements of resources in a network; work is still in progress on this.

A new version of [FLBB] was prepared and submitted for publication.

### F. Atomic Registers

Vitanyi and Awerbuch [VA] have studied the feasibility of atomic shared register access by asynchronous hardware. The problem is to construct multivalued registers which can be read and written asynchronously by many processes in a consistent fashion. Moreover, it is required that any process should be able to proceed without waiting for any other process. Using atomic 1-reader, 1-writer registers, they have constructed atomic multireader, multiwriter registers using unbounded tags.

In [B], we give an algorithm allowing two processors with one-writer atomic registers with 2n values each to simulate a n-value atomic register that they can both write. This is still work in progress.

### G. Other Network Problems

[AG,AM,ACMG] describe new results on some interesting network problems. [AG] gives an efficient, though complicated, new algorithm for performing breadth-first search of a distributed network. [AM] involves development of a new algorithm for detecting and breaking deadlocks among processes in a network.

[AGMS] gives a new algorithm for carrying out a "global coin toss" in an unreliable network. This is a very fundamental problem, since many existing protocols rely on the existence of such a coin. They have proposed a new, efficient protocol that produces a provably fair coin in the presence of malicious adversaries. Their solution uses weak cryptographic assumptions.

Paul Vitanyi has been working on a problem of distributed control [KV]. He has studied the number of messages required for matching pairs of mobile processes in a multiprocessor network; this is a measure for the cost of setting up temporary communication between such processes. He has established lower bounds on the average number of point-to-point transmissions between any pair of nodes in this context. Applications of the results include lower bounds on the number of messages required to implement a distributed name-server, and to solve distributed mutual exclusion and distributed resource allocation problems.

Coan has worked on limitations on database availability when networks partition [CoOK]. In designing fault-tolerant distributed database systems, a frequent goal is to make the system highly available despite component failure. They describe a way of measuring availability and prove a lower bound on the availability that can be achieved by any on-line replicated data management protocol that maintains database consistency. This bound holds under a certain uniformity assumption on the pattern of data accesses by transactions.

## II. Models

Gene Stark's PhD thesis [S1] was completed during this reporting period; it contains a formal foundation for a theory of specification of modules in distributed systems.

Mark Tuttle has been working with Nancy Lynch [LT] on resource allocation algorithms and their correctness proofs. We have been using a new "levels of abstraction" organization for proofs of correctness of certain distributed algorithms. In particular, we have been applying it to prove the correctness of a new design for a distributed arbiter algorithm. The proof organization provides a new way of understanding distributed algorithms in terms of the "abstract knowledge" present at each node.

In order to carry out a clean, hierarchical proof, we have found it necessary to develop a clear formal foundation for this work. A basic semantic model for concurrent computation has been defined; it is based on a simple component which we call an I/O automaton. One important aspect of this model is the division of process actions into input actions and output actions, which permits us to model the notion of a "fair computation" easily. Our model captures the game-theoretic nature of distributed computation. It includes treatment of both finite and infinite properties of module behavior. It allows organization of algorithms using several conceptual levels of abstraction.

The I/O automaton model has been applied to several different areas of concurrent computing. For example, Jennifer Lundelius Welch is using the model to describe shared memory algorithms. In particular, she is describing a well-known n-process mutual exclusion algorithm of Peterson and Fischer in a more modular way than they do. Two advantages are gained. First, any 2-process mutual exclusion algorithm can be used as a subroutine in the tournament tree in her formulation, instead of just Peterson and Fischer's 2-process solution. Second, the time performance can be reduced from $O(n^2)$ to $O(n \log n)$. An important aspect of this work is the development (still in progress) of time measures for asynchronous systems, to be integrated with the I/O automaton model.

Another application of the model has been to produce a more modular description of a family of solutions to Chandy and Misra's Drinking Philosophers problem. Unlike their original presentation, the new description produces a drinking philosophers algorithm from an arbitrary dining philosophers algorithm as a subroutine. This work is at preliminary stages; it has not yet been written up.

Also, with Dr. Leslie Lamport, I have been attempting a formal proof, in levels of abstraction, of the very well-known minimum spanning tree algorithm of Gallager, Humblet and Spira. This has met with only partial success so far.

The principal contribution of [DS] is the introduction of a new type of reduction designed expressly for distributed systems. This reduction classifies distributed problems by the communication requirements of their solutions.

In [KS], we propose a new method for the analysis of cooperative and antagonistic properties of communicating finite state processes (FSP's). This algebraic technique is based on a composition operator and the notion of "possibility equivalence" among FSP's. We demonstrate its utility by showing that potential blocking, lockout, and termination can be efficiently decided for loosely connected networks of tree FSP's. If not all acyclic FSP's are trees, then the cooperative properties become NP-complete and the antagonistic ones PSPACE-complete. We also have related results for tightly coupled networks and for the considerably harder cyclic process case.

Lundelius has shown [L2] how a distributed system with synchronous processors and asynchronous communication can be simulated by a system in which both processors and communication are asynchronous, in the presence of various types of processor failures. Consequently, a result of Dolev, Dwork and Stockmeyer, that no consensus protocol in a system with synchronous processors and asynchronous communication can tolerate even one failstop processor, follows from the result of Fischer, Lynch and Paterson, that fault-tolerant consensus is impossible when both processors and communication are asynchronous.

## III. Concurrency Control and Recovery

### A. Nested Transactions

We have been engaged in an ambitious project to provide a natural formal foundation for concurrency control and resiliency. Our goal is to provide a framework within which researchers, developers and implementers can discuss interesting requirements and algorithms for distributed transaction-processing systems. This area is of critical importance to distributed computing, but the work is currently described in hundreds of unrelated research papers, with no common framework to aid in comprehension. We are especially interested in a theory to underlie "nested transactions", an important new language construct for distributed computing.

The paper [L1], on a preliminary model for nested transactions and a proof of correctness for an exclusive locking algorithm, was revised and sent back to Advances in Computing Research for final publication.

The paper [LM] contains the first reports on our results on a new, cleaner and more expressive model for nested transaction concurrency control and recovery. This framework appears to be satisfactory for all its purposes. The paper includes a statement of a basic correctness condition to be satisfied by all nested transaction systems. It also contains a correctness proof for an exclusive locking algorithm. Part of this effort includes description of the implementation of data objects with resiliency properties, in terms of basic data objects without such properties. The presentation and proofs are much simpler and give more insight than previous work.

We have begun to extend this work in several different directions. With Alan Fekete, Michael Merritt and Prof. Bill Weihl, we have proved correctness of an important practical algorithm, Moss' read-write locking algorithm for nested transactions [FLMW]. This algorithm is currently implemented in the ARGUS system.

With Prof. Maurice Herlihy and Merritt and Weihl [HLMW], I have been working on describing and proving correctness of several algorithms for the detection and elimination of "orphan" transactions -- transactions with ancestors that abort. If not managed properly, orphan transactions pose a danger of causing damage, or wasting system resources, so several algorithms have been designed for managing orphans in various systems. What has not been clear until now is exactly why these algorithms are correct, or even what it means for them to be correct. We have been able to describe precise correctness conditions within our model, and have shown correctness of two important orphan-management algorithms. The proofs have been very clear, easy and short, in marked contrast to earlier attempts to carry out such proofs. We are currently studying some other orphan-management algorithms, including some that work in the presence of system node crashes which lose the contents of volatile memory.

With Ken Goldman, I have been working on modelling replicated data management algorithms [GL]. Here, we are interested in algorithms for managing replicated data in the presence of site and communication failures (including network partitions). This work involves unifying the known results for non-nested transactions) and extending the results to nested transactions. All of this work has proceeded very successfully, and papers are in various advanced stages of progress.

In somewhat less advanced stages of progress is work which is aimed toward a general theorem about nested transactions and abstract objects, work on modelling crashes in distributed networks, and work on timestamp-based algorithms for implementing nested transactions. ALl of these are important areas, and we will continue this work in the future.

### B. Highly Available Replicated Data Systems

Nancy Lynch's consulting work at CCA has led to several new research ideas. CCA is building a distributed transaction processing system (SHARD) which is intended to work in a SAC environment, in which communication is very unreliable. I have been involved in the system's design and specification. In particular, I have helped design reliable broadcast algorithms for use with unreliable packet radio communications [GLBKSS], and algorithms to change system configuration during its operation [SL].

Most recently, I have been developing a set of correctness properties to describe the guarantees which SHARD is able to make to its users [LBS]. Systems such as SHARD sacrifice strong correctness conditions such as "serializability", in the interests of performance. In environments with unreliable communication, it may be necessary to do this. It is important, however, to be able to make some precise guarantees about what such systems can do. The guarantees made by the system include nonstop operation, preservation of data consistency in case of nonfaulty communications, bounds of "costs" of inconsistencies during certain kinds of faulty communications, and certain "fairness" guarantees. It is important to make such guarantees explicit, so that the very novel approach embodied in the CCA system can be compared objectively with more traditional approaches

### c. Publications and Technical Reports

[AM] Awerbuch, B., and Micali, S., "Complexity of Resolution and Detection of Deadlocks," *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[ACMG] Awerbuch, B., Chor, B., Micali, S., and Goldwasser, S., "Verifiable Secret Sharing and Achieving Simultaneity in the Presence of Faults," *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[AG] Awerbuch, B., and Gallager, R., "Distributed Breadth-First-Search Algorithms," *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[B] Bloom, B. "Building Multiple-Writer Atomic Registers from Single-Writer Atomic Registers," Work in progress.

[BL] Burns, J.E., and Lynch, N.A., "The Byzantine Firing Squad Problem," TM-275, Laboratory for Computer Science, M.I.T., April 1985, also to appear in *Advances in Computing Research*.

[C1] Coan, B., "Communication-Efficient Canonical Forms for Fault-Tolerant Distributed Protocols," *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada, (August 11-13, 1986), pp. 63-72.

[CC] Chor, B., and Coan, B.A., "A Simple and Efficient Randomized Byzantine Agreement Algorithm," *IEEE Transactions on Software Engineering*, Vol. SE-11, No. 6, pp. 531-539, June 1985. Also, *Proceedings 4th Symposium on Reliability in Distributed Software and Database Systems*, Sheraton Inn, Northwest Washington, Silver Spring, MD., (October 15-17, 1984), pp. 98-106. Also, TM-266, Laboratory for Computer Science, M.I.T., August 1984.

[CD] Coan, B.A., and Dwork, C., "Simultaneity is Harder than Agreement," *Proceedings 5th Symposium on Reliability in Distributed Software and Database Systems*, Marriott Hotel, Los Angeles, CA., (January 13-15, 1986), pp. 141-150.

[CDDS] Coan, B., Dolev, D., Dwork, C., and Stockmeyer, L., "The Distributed Firing Squad Problem," *Proceedings of the 17th Annual ACM Symposium on Theory of Computing*, Providence, R.I., (May 6-8, 1985), pp. 335-345.

[CL] Coan, B., and Lundelius, J. "Transaction Commit in a Realistic Fault Model," *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada (August 11-13, 1986), pp. 40-51.

[CoOK] Coan, B., Oki, B. M. and Kolodner, E. K., "Limitations on Database Availability when Networks Partition," *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada, (August 11-13, 1986), pp. 187-195.

[CMS] Chor, B., Merritt, M. and Shmoys, D. B., "Simple Constant-Time Consensus Protocols in Realistic Failure Models," *Proceeding of the Fourth Annual ACM Symposium on Principles of Distributed Computing*, Minaki, Ontario, Canada, (August 5-7, 1985), pp. 152-162.

[DLPSW] Dolev, D., Lynch, N.A., Pinter, S. S., Stark, E. W., and Weihl, W. E., "Reaching Approximate Agreement in the Presei.ce of Faults," *Proceedings of 3rd Annual IEEE Symposium on Reliability in Distributed Software and Database Systems*, Clearwater, FL., (October 17-19, 1983) pp. 145-154; also TM-276, Laboratory for Computer Science, MIT, May 1985, also to appear in *Journal of the Association for Computing Machinery*.

[DLS] Dwork, C., Lynch, N., and Stockmeyer, L. "Consensus in the Presence of Partial Synchrony," *Proceedings of 3rd ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, Vancouver, B.C., Canada (August 27-29, 1984), pp. 103-118; also TM-270, Laboratory for Computer Science, MIT, October 1984, and TM-270 Laboratory for Computer Science, MIT, October 1985, [revision of October, 1984 version of this paper]. To appear in *JACM*.

[DM] Dwork, C., and Moses, Y., "Knowledge and Common Knowledge in Byzantine Environments I: Crash Failures (Preliminary Ver.)," *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*, Monterey, CA., (March 19-22, 1986), pp. 149-170.

[DS] Dwork, C., and Skeen, D., "Patterns of Communication in Consensus Protocols," *Proceedings of Third Annual ACM Symposium on Principles of Distributed Computing*, Vancouver, B.C., Canada, (August 27-29, 1984), pp. 143-153.

[FGGL] Fischer, Griffeth, Guibas, Lynch Fischer, M.J., Griffeth, N.D., Guibas, L.J., and Lynch, N., "Optimal Placement of Identical Resources in a Tree," To appear in *Information and Control*.

[FL] Frederickson, G.N., and Lynch, N.A., "Electing a Leader in A Synchronous Ring," TM-277 [Revision of MIT/LCS/TM-259 and March 1985 version of this paper], Laboratory for Computer Science, MIT, July 1985. To appear in *JACM*.

[FLBB] Fischer, M., Lynch, N. A., Burns, J., and Borodin, A., "The Colored Ticket Algorithm," TM-269, Laboratory for Computer Science, MIT, August 1983.

[FLM] Fischer, M.J., Lynch, N.A., and Merritt, M., "Easy Impossibility Proofs for Distributed Consensus Problems," *Proceedings of the Fourth Annual Symposium on Principles of Distributed Computing*, Minaki, Ontario, Canada, (August 5-7, 1985) pp. 59-70; also TM-279, Laboratory for Computer Science, M.I.T., June 1985, also in inaugural issue of *Distributed Computing* 1, 1 (1986), pp. 26-39.

[FLMW] Fekete, A. D., Lynch, N., Merritt, M., and Weihl, W., "Nested Transactions and Read/Write Locking," *Proc. 6th ACM Symp. on Principles of Database Systems*, San Diego, California, March 1987, to appear.

[GL] Goldman, K., and Lynch, N., "Data Replication in Nested Transaction Systems," in progress.

[GLBKSS] Garcia-Molina, H., Lynch, N., Blaustein, B., Kaufman, C., Sarin, S., and Shmueli, O., "Notes on a Reliable Broadcast Protocol," Internal CCA report.

[HLMW] Herlihy, M., Lynch, N., Merritt, M., and Weihl, W., On the Correctness of Orphan Elimination Algorithms, Submitted for publication.

[KS] Kanellakis, P.C., and Smolka, S.A., "On the Analysis of Cooperation and Antagonism in Networks of Communicating Processes," *Proceedings of the Fourth Annual ACM Symposium on Principles of Distributed Computing,* Minaki, Ontario, Canada, (August 5-7, 1985), pp. 23-38.

[KV] Kranakis, E., and Vitanyi, P.M.B., "Distributed Control in Computer Networks and Cross-Sections of Multidimensional Bodies," MIT/LCS/TM-304, March 1986. (Submitted to *Journal of the ACM*).

[L1] Lynch, N.A., "Concurrency Control for Resilient Nested Transactions," *Proceedings of 2nd ACM SIGACT-SIGMOD Symposium on Principles of Database Systems,* Atlanta, Ga., (March 21-23, 1983), pp. 166-181; also TR-285 Laboratory for Computer Science, MIT, February 1983, also in *Advances in Computing Research,* 3 (1986), pp. 335-373.

[LBS] Lynch, N., Blaustein, B., and Siegel, M., "Correctness Conditions for Highly Available Replicated Databases," *Proceedings of 5th ACM SIGACT-SIGOPS Symposium on Principles of Distributed* Computing), (August 1986) pp. 11-28.

[Lu1] Lundelius, J., "Synchronizing Clocks in a Distributed System,", M.S. Thesis, TR-335, Laboratory for Computer Science, MIT, August 1984.

[Lu2] Lundelius, J.., "Simulating Synchronous Processors," to appear in *Information and Control.*

[LuL1] Lundelius, J. and Lynch, N., "A New Fault-Tolerant Algorithm for Clock Synchronization," *Proceedings of 3rd ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing,* Vancouver, B.C., Canada (August 27-29, 1984) pp. 75-88, also TM-265, Laboratory for Computer Science, MIT, July 1984. To appear in *Information and Control.*

[LuL2] Lundelius, J., and Lynch, N., "An Upper and Lower Bound for Clock Synchronization," *Information and Control* 62, 2/3 (August-September 1984), pp. 190-204.

[LGFG] Lynch, N.A., Griffeth, N., Fischer, M., and Guibas, L., "Probabilistic Analysis of a Network Resource Allocation Algorithm," *AMS Workshop on Probabilistic Algorithms (ABSTRACT ONLY)* (June 1982). Also, in *Information and Control* 68, 1-3 (January-February-March 1986); also TM-278, Laboratory for Computer Science, MIT, June 1985.

[LM] Lynch, N.A., and Merritt, M., "The Theory of Nested Transactions: Concurrency Control and Resiliency," *(ICDT'86) International Conference on Database Theory,* Rome, Italy (September 8-10, 1986). Also, TR-367, Laboratory for Computer Science, MIT, July 1986. Submitted for publication.

[LT] Lynch, N., and Tuttle, M., "Correctness Proofs for Distributed Algorithms," in progress.

[MT] Moses, Y. and Tuttle, M., "Common Knowledge and Simultaneous Actions in the Presence of Failures". Submitted for publication. To Appear as an MIT Technical Report.

[S1] Stark, E., "Foundations of a Theory of Specification for Distributed Systems," Ph.D Thesis, TR-342, Laboratory for Computer Science, MIT, August, 1984.

[SL] Sarin, S., and Lynch, N.A., "Discarding Obsolete Information in a Replicated Database System," To appear in *IEEE Transactions on Software Engineering,* (December 1986).

[VA] Vitanyi, P.M.B., and Awerbuch, B., "Atomic Shared Register Access by Asynchronous Hardware", *27th Annual IEEE Symposium on Theory of Computing,* Toronto, 1986.

### d. Participating Scientific Personnel

Baruch Awerbuch, Postdoctoral Research Associate
Bard Bloom, Graduate Student
James Burns, Postdoctoral Research Associate
Chris Clifton, Graduate Student, MS, June 1986
Cynthia Dwork, Postdoctoral Research Associate
Alan Fekete, Graduate Student
Kenneth Goldman, Graduate Student
John Goree, Graduate Student, MS, January 1983
Paris Kanellakis, Postdoctoral Research Associate
Jennifer Lundelius, Graduate Student, MS, August 1984
Everett McKay, Graduate Student, BS/MS, January 1985
Michael Merritt, Postdoctoral Research Associate
Yoram Moses, Postdoctoral Research Associate
Neil Savasta, Graduate Student, SB, June 1984
Eugene Stark, Graduate Student, Ph.D, August 1984
Mark Tuttle, Graduate Student
Paul Vitanyi, Postdoctoral Research Associate
Shmuel Zaks, Postdoctoral Research Associate

## e. Bibliography

[AM] Awerbuch, B., and Micali, S., "Complexity of Resolution and Detection of Deadlocks." *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[ACMG] Awerbuch, B., Chor, B., Micali, S., and Goldwasser, S., "Verifiable Secret Sharing and Achieving Simultaneity in the Presence of Faults." *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[AG] Awerbuch, B., and Gallager, R., "Distributed Breadth-First-Search Algorithms." *IEEE Symposium on Foundations of Computer Science*, October 1985, Portland, Oregon.

[B] Bloom, B. "Building Multiple-Writer Atomic Registers from Single-Writer Atomic Registers," Work in progress.

[BL] Burns, J.E., and Lynch, N.A., "The Byzantine Firing Squad Problem," TM-275, Laboratory for Computer Science, M.I.T., April 1985, also to appear in *Advances in Computing Research*.

[C1] Coan, B., "Communication-Efficient Canonical Forms for Fault-Tolerant Distributed Protocols," *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada, (August 11-13, 1986), pp 63-72.

[CC] Chor, B., and Coan, B.A., "A Simple and Efficient Randomized Byzantine Agreement Algorithm," *IEEE Transactions on Software Engineering*, Vol. SE-11, No. 6, pp. 531-539, June 1985. Also, *Proceedings 4th Symposium on Reliability in Distributed Software and Database Systems*, Sheraton Inn, Northwest Washington, Silver Spring, MD., (October 15-17, 1984), pp. 98-106. Also, TM-266, Laboratory for Computer Science, M.I.T., August 1984.

[CD] Coan, B.A., and Dwork, C., "Simultaneity is Harder than Agreement," *Proceedings 5th Symposium on Reliability in Distributed Software and Database Systems*, Marriott Hotel, Los Angeles, CA., (January 13-15, 1986), pp. 141-150.

[CDDS] Coan, B., Dolev, D., Dwork, C., and Stockmeyer, L., "The Distributed Firing Squad Problem," *Proceedings of the 17th Annual ACM Symposium on Theory of Computing*, Providence, R.I., (May 6-8, 1985), pp. 335-345.

[CL] Coan, B., and Lundelius, J. "Transaction Commit in a Realistic Fault Model." *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada (August 11-13, 1986), pp. 40-51.

[CoOK] Coan, B., Oki, B. M. and Kolodner, E. K., "Limitations on Database Availability when Networks Partition," *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, Calgary, Alberta, Canada, (August 11-13, 1986), pp. 187-195.

[CMS] Chor, B., Merritt, M. and Shmoys, D. B., "Simple Constant-Time Consensus Protocols in Realistic Failure Models," *Proceeding of the Fourth Annual ACM Symposium on Principles of Distributed Computing*, Minaki, Ontario, Canada, (August 5-7, 1985).

pp. 152-162.

[DLPSW] Dolev, D., Lynch, N.A., Pinter, S. S., Stark, E. W., and Weihl, W. E., "Reaching Approximate Agreement in the Presence of Faults," *Proceedings of 3rd Annual IEEE Symposium on Reliability in Distributed Software and Database Systems*, Clearwater, FL., (October 17-19, 1983) pp. 145-154; also TM-276, Laboratory for Computer Science, MIT, May 1985, also to appear in *Journal of the Association for Computing Machinery*.

[DLS] Dwork, C., Lynch, N., and Stockmeyer, L. "Consensus in the Presence of Partial Synchrony," *Proceedings of 3rd ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, Vancouver, B.C., Canada (August 27-29, 1984), pp. 103-118; also TM-270, Laboratory for Computer Science, MIT, October 1984, and TM-270 Laboratory for Computer Science, MIT, October 1985, [revision of October, 1984 version of this paper]. To appear in *JACM*.

[DM] Dwork, C., and Moses, Y., "Knowledge and Common Knowledge in Byzantine Environments I: Crash Failures (Preliminary Ver.)," *Proceedings of the Conference on Theoretical Aspects of Reasoning About Knowledge*, Monterey, CA., (March 19-22, 1986), pp. 149-170.

[DS] Dwork, C., and Skeen, D., "Patterns of Communication in Consensus Protocols," *Proceedings of Third Annual ACM Symposium on Principles of Distributed Computing*, Vancouver, B.C., Canada, (August 27-29, 1984), pp. 143-153.

[FGGL] Fischer, Griffeth, Guibas, Lynch Fischer, M.J., Griffeth, N.D., Guibas, L.J., and Lynch, N., "Optimal Placement of Identical Resources in a Tree," To appear in *Information and Control*.

[FL] Frederickson, G.N., and Lynch, N.A., "Electing a Leader in A Synchronous Ring," TM-277 [Revision of MIT/LCS/TM-259 and March 1985 version of this paper], Laboratory for Computer Science, MIT, July 1985. To appear in *JACM*.

[FLBB] Fischer, M., Lynch, N. A., Burns, J., and Borodin, A., "The Colored Ticket Algorithm," TM-269, Laboratory for Computer Science, MIT, August 1983.

[FLM] Fischer, M.J., Lynch, N.A., and Merritt, M., "Easy Impossibility Proofs for Distributed Consensus Problems," *Proceedings of the Fourth Annual Symposium on Principles of Distributed Computing*, Minaki, Ontario, Canada, (August 5-7, 1985) pp. 59-70; also TM-279, Laboratory for Computer Science, M.I.T., June 1985, also in inaugural issue of *Distributed Computing* 1, 1 (1986), pp. 26-39.

[FLMW] Fekete, A. D., Lynch, N., Merritt, M., and Weihl, W., "Nested Transactions and Read Write Locking," *Proc. 6th ACM Symp. on Principles of Database Systems*, San Diego, California, March 1987, to appear.

[GL] Goldman, K., and Lynch, N., "Data Replication in Nested Transaction Systems," in progress.

[GLBKSS] Garcia-Molina, H., Lynch, N., Blaustein, B., Kaufman, C., Sarin, S., and Shmueli, O., "Notes on a Reliable Broadcast Protocol," Internal CCA report.

[HLMW] Herlihy, M., Lynch, N., Merritt, M., and Weihl, W., On the Correctness of Orphan Elimination Algorithms, Submitted for publication.

[KS] Kanellakis, P.C., and Smolka, S.A., "On the Analysis of Cooperation and Antagonism in Networks of Communicating Processes," *Proceedings of the Fourth Annual ACM Symposium on Principles of Distributed Computing,* Minaki, Ontario, Canada, (August 5-7, 1985), pp. 23-38.

[KV] Kranakis, E., and Vitanyi, P.M.B., "Distributed Control in Computer Networks and Cross-Sections of Multidimensional Bodies," MIT/LCS/TM-304, March 1986. (Submitted to *Journal of the ACM*).

[L1] Lynch, N.A., "Concurrency Control for Resilient Nested Transactions," *Proceedings of 2nd ACM SIGACT-SIGMOD Symposium on Principles of Database Systems,* Atlanta, Ga., (March 21-23, 1983), pp. 166-181; also TR-285 Laboratory for Computer Science, MIT, February 1983, also in *Advances in Computing Research,* 3 (1986), pp. 335-373.

[LBS] Lynch, N., Blaustein, B., and Siegel, M., "Correctness Conditions for Highly Available Replicated Databases," *Proceedings of 5th ACM SIGACT-SIGOPS Symposium on Principles of Distributed* Computing), (August 1986) pp. 11-28.

[Lu1] Lundelius, J., "Synchronizing Clocks in a Distributed System,", M.S. Thesis, TR-335, Laboratory for Computer Science, MIT, August 1984.

[Lu2] Lundelius, J., "Simulating Synchronous Processors," to appear in *Information and Control.*

[LuL1] Lundelius, J. and Lynch, N., "A New Fault-Tolerant Algorithm for Clock Synchronization," *Proceedings of 3rd ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing,* Vancouver, B.C., Canada (August 27-29, 1984) pp. 75-88, also TM-265, Laboratory for Computer Science, MIT, July 1984. To appear in *Information and Control.*

[LuL2] Lundelius, J., and Lynch, N., "An Upper and Lower Bound for Clock Synchronization," *Information and Control* 62, 2/3 (August-September 1984), pp. 190-204.

[LGFG] Lynch, N.A., Griffeth, N., Fischer, M., and Guibas, L., "Probabilistic Analysis of a Network Resource Allocation Algorithm," *AMS Workshop on Probabilistic Algorithms (ABSTRACT ONLY)* (June 1982). Also, in *Information and Control* 68, 1-3 (January-February-March 1986); also TM-278, Laboratory for Computer Science, MIT, June 1985.

[LM] Lynch, N.A., and Merritt, M., "The Theory of Nested Transactions: Concurrency Control and Resiliency," *(ICDT'86) International Conference on Database Theory,* Rome, Italy (September 8-10, 1986). Also, TR-367, Laboratory for Computer Science, MIT, July 1986. Submitted for publication.

[LT] Lynch, N., and Tuttle, M., "Correctness Proofs for Distributed Algorithms," in progress.

[MT] Moses, Y. and Tuttle, M., "Common Knowledge and Simultaneous Actions in the Presence of Failures". Submitted for publication. To Appear as an MIT Technical Report.

[S1] Stark, E., "Foundations of a Theory of Specification for Distributed Systems," Ph.D Thesis, TR-342, Laboratory for Computer Science, MIT, August, 1984.

[SL] Sarin, S., and Lynch, N.A., "Discarding Obsolete Information in a Replicated Database System," To appear in *IEEE Transactions on Software Engineering*, (December 1986).

[VA] Vitanyi, P.M.B., and Awerbuch, B., "Atomic Shared Register Access by Asynchronous Hardware", *27th Annual IEEE Symposium on Theory of Computing*, Toronto, 1986.

END

3-87

DTIC